# A FRAMEWORK FOR ALIGNING AI SYSTEMS WITH HUMAN VALUES

A
Synopsis
Submitted in partial fulfillment of the requirements of

**Doctor of Philosophy**

**In**

**Physics and Computer Science**

Submitted by

**Sukrati Chaturvedi**

Under the Supervision of

**Prof. C. Vasantha Lakshmi**
**Supervisor**
Department of Physics & Computer Science
Faculty of Science

**Prof. C. Patvardhan**
**Co-Supervisor**
Department of Electrical Engineering
Faculty of Engineering

**DEPARTMENT OF PHYSICS & COMPUTER SCIENCE**
**FACULTY OF SCIENCE**
**DAYALBAGH EDUCATIONAL INSTITUTE (DEEMED UNIVERSITY)**
**DAYALBAGH, AGRA (UP) – 282005**
**2020**

## Introduction

### 1.1 Artificial Intelligence

Artificial Intelligence (AI) is the field of making computers perform tasks that, when performed by humans, require intelligence. Much success has been achieved in teaching computers various tasks viz., playing games like Chess and Go, perception etc. Significant progress has been made in this field from mastering the games to self-driving vehicles. Alpha Go [1] became the first program to defeat the world champion in the game of GO. The next version, Alpha Go Zero [2] was its own teacher which defeated Alpha Go easily. No prior knowledge was given to it but the basic rules of game only. As a result, it became the best Go player in the world without any human intervention and using no historical data, solely from self-play using Reinforcement Learning. These are dedicated to particular tasks and can only perform those tasks in a narrow domain. Hence, this kind of AI is referred to as Artificial Narrow Intelligence.

Recent attempts have focused on creating AI that can act in more than one domain. Accordingly, AI is said to be of three kinds: Artificial Narrow Intelligence (ANI) as referred above, Artificial General Intelligence (AGI), and Artificial Super Intelligence (ASI) [3]. These can be distinguished as follows.

(i) ANI: This is the first level of AI that specializes in one area only. For example- AI system Deep Blue [4] that can beat the world chess champion, but that's the only task it can perform. Appearance of this level of AI development is clearly visible in technologies that we use every day. These advancements will increase with every coming year and technology will become more complex, addictive and pervasive as it becomes more intelligent.

(ii) AGI: This is the second level of AI that can perform any intellectual task that a human can perform, meaning it has ability of planning, reasoning, solving problem, learning from experiences [5] etc.

(iii) ASI: Further progress will lead us towards a new level of AI – Artificial Super Intelligence (ASI), which will surpass human intelligence and be practically much smarter than best human brain in every field [6].

Currently, the lowest level of AI i.e. ANI has been achieved in many domains. Examples include Google search, email spam filters etc. The current efforts that have heralded the AI revolution is the path from ANI, through AGI, to ASI. According to some researchers, these advancements in AI indicate that AGI is achievable in next five to ten years although there are some who are more conservative in their predictions [7].

A lot of debate has been generated about the developments and implications of the impending AGI andASI among the researchers in recent times. Some opinions are as follows.

- According to Robin Hanson[8], within a century, computers will be as capable as brain and such computers will be cheaper than brain and will therefore displace the brains in almost all job tasks.
- Martine Rothblatt[9] says that in 21st century we will have computers as capable as brain.
- According to Ray Kurzweil[10], computers will exceed human capabilities by 2029. He adds that AI represents itself as a brain extender.
- Nick Bostrom[11] predicts that we could have human level AI within a small number of decades though there is a great uncertainty on both the lower and upper sides of the estimates.
- According to Rodney Brook[12], computers will be as intelligent as humans in may be many hundred years but will be as intelligent and as conscious as dogs in next 50 to 100 years.
- Garry Marcus[13] states that computers are already far more capable than human brains in many fields like arithmetic and memory etc. but it could still take 20 to 50 years more to have the ability to read and comprehend and reason about novel situations as fluently as people can.
- JurgenSchmidhuber[14] predicts that "soon" AI will be as smart as human brains as every five years computing is getting roughly ten times cheaper.

| **Researcher** | **Prediction about evolution of ASI** |
|---|---|
| Robbin Hanson | Within a century |
| Martin Rothblatt | In 21st century |
| Ray Kurzweil | By 2029 |

| | |
|---|---|
| Nick Bostrom | Within a smaller number of decades |
| Rodney Brook | Not in this lifetime but in many hundred years |
| Garry Marcus | 20 to 50 years |
| JurgenSchmidhuber | Soon |

Google announced the prototype of a self-driven car [15] (the Google Car) in the year 2010. According to the executives of Google at the time, the goal of the Google Car was to save people's time, to minimize the traffic accidents by preventing them and to reduce the carbon emissions. A self-driven vehicle would not require any human involvement and it would operate without any human control [16]. Modern self-deriving vehicles are able to sense their local environment and classify different types of objects that they detect. Also, self-driven vehicles can identify appropriate navigation paths while following transportation rules by interpreting sensory information [17].

The timeline is getting closer. Driverless cars were not at all feasible even a few decades back but are now regarded as inevitable. Not only driverless cars but also the applications like Ola and Uber were not visualized some time back. Today a cab can be booked through the app and the application would get our location and the driver would be at our place in just a few minutes. Perception is being widely regarded as a solved problem with applications being launched on a daily basis. All these developments have given major boost to the belief that AGI, and perhaps even ASI, would soon be a reality.

### 1.2 Major approaches for developing AI Systems

AI has gone through several phases of transformation with the approaches at any given point of time reflecting the understanding of intelligence and the available technology for implementing it. Initially, it was hypothesized that an intelligent system can be created by some mechanism for representing knowledge and some mechanisms for implementing reasoning using this knowledge. Typically, if-else rules were used to this end. These approaches did work in the beginning and some small systems with severe domain restrictions were developed to demonstrate their workability. But, it was soon realized that, these systems were not scalable and were extremely brittle. A soon as one tried to enhance them by increasing their domain of operation they would fail.

Alternative approaches that used learning from examples as the main paradigm of constructing AI systems were then developed. These were called Neural Networks as they derived motivation from

human brain. These were also found to be limited in their applicability as the returns tapered off as described in economics by the "Law of diminishing Returns" when attempts were made to use larger data sets towards attaining higher accuracies with better generalization. Availability of validated ground data and large computational firepower were the two major constraints in achieving further success.

The scenario changed with the availability of faster machines including CUDA machines and clusters of CUDA machines. Further, availability of huge data sets became more ubiquitous. Recent efforts in developing AI systems have focused on utilizing Machine Learning wherein the system is trained with enormous amounts of ground truth data to enable it to perform the intended task. In these endeavors the most important component is the specially prepared and validated ground truth data that is utilized to train the system. The quality of performance of the AI system depends heavily on the quality and quantity of the ground truth data. If there is bias in the data, it will get reflected in the Machine Learning system that is created using that data. Further, any noise in the data would result in performance degradation of the system. Therefore, it is important not to work with such data in developing Machine Learning systems. Identifying the noisy elements and cleaning up the data requires tremendous amount of effort. Preparation of the right quality and quantity of data is, therefore, critical for the success of the Machine Learning AI endeavors.

Some approaches that have recently gained prominence for the development of AI systems are as follows.

### 1.2.1 Deep learning

Deep learning (DL) is a branch of machine learning (ML) research, whose goal is to draw ML nearer its objective, AI [18]. DL algorithms are used to train deep networks with large data sets [19]. Deep learning or a deep neural network has a succession of multiple processing layers. Each layer consists of a non-linear transformation and the sequence of these transformations leads to learning difficult level of abstractions. DL extracts complex features from high dimensional data and applies them to develop a model that relates inputs to outputs. And this is exactly how our mind works by training the data it captures or sees. Therefore, we can predict that DL might work as mind for AI.

DL is solving problems that were not solved by AI community even after the best attempts. The key aspect of DL is that human engineers don't design layers of features; they are learned from data using a general purpose algorithm [20].

DL has lot of applications like speech recognition, image processing, language processing etc. Zhanyi Wang has presented a traffic identification system using deep learning [21]. The field is growing quickly with new techniques reported continuously. DL methods achieve high accuracy for tasks where a large amount of data is available.

### 1.2.2 Reinforcement learning

Reinforcement Learning (RL) is an area of AI, which has emerged as an effective tool towards building AI systems and solving decision making problems. RL has achieved impressive success in the recent years, it is now able to play and win various games [22] like Go and Atari etc. Further, RL combined with Markov Decision Process (MDP) can model a self-driving vehicle in an environment comprising of road with obstacles and desired goals [23].

Main goal of RL is to maximize the rewards by interacting with an environment, taking different actions and experiencing many failures and successes [24]. The agent is not guided to take a predefined action. RL is same as natural learning process where the teacher or guide is not present and the student learns through trial and error method.

The main components of RL system are: policy, reward signal, value function and model [25][26].

- The **policy ($\pi$)** guides an agent on how to behave under certain circumstances. The policy maps states into actions. Evolving a best policy is the ultimate goal of a RL system.
- The **reward signal (R)** indicates how good or bad an event is. Also, it defines the goal of the problem where main objective of the agent is to maximize the total reward. Reward is the main factor for updating policy.
- The **value function** is prediction of total future rewards which is used to evaluate the states and select between different actions accordingly.
- **Model** is an optional component of RL. It allows the predictions to be made about the behavior of the environment.

### 1.2.3 Deep reinforcement learning

Deep reinforcement learning (DRL) is a combination of deep learning and reinforcement learning techniques. DRL has been able to solve a number of decision-making problems that were previously unsolved [27]. Reasonably good AI systems have emerged using this combination. In fact, the main contributor of AlphaGo David Silver,mentioned that [28]

Artificial Intelligence= Reinforcement learning + Deep learning

DRL bridges the gap between sensed input and the response action resulting in an AI ready toface any challenge. Significant progress has been made by combining advances in DL with RL, resulting in the deep Q network, which is capable of human level performance on many video games [29].

### 1.2.4 Evolutionary algorithms

Evolutionary algorithms (EA) are inspired by biological evolutions such as reproduction, mutation etc. EA works in four steps: initialization, selection, reproduction and termination. The process is started by generating random solution population initially, and then the fitness of all the individuals is calculated. After that, best fit individuals are selected for reproduction. Those selected individuals are bred through crossover and mutation operation to give birth to offspring. Then fitness of new individual is calculated and then the least fit individual of parent generation is replaced by the new individual. All this process is repeated till we get the optimal solution.

Wilson et al. [30] proposed playing Atari using evolutionary algorithms whereas Hausknecht el al. [31] introduced an approach for playing Atari games using neuro-evolution. EAs could also be used in solving multi-objective problems (i.e., problem having several conflicting objects) with a set of optimal solutions. Zitzler et al. [32] presented comparison between different evolutionary approaches for multi-objective optimization using six different test functions. And the suggested test functions provided sufficient complexity for comparing different multi-objective optimizers.

### 1.2.5 Natural language processing

Natural Language Processing (NLP) techniques are required whenever we want AI to perform any task as per our instructions given in natural language. For example- Alexa [33] responds to instructions given in natural language and performs tasks such as switching off the lights or playing some song as per the instructions.

NLP forms the basis of many applications such as Question AnswerSystems(QAS), Automated story generation Systems, text summarization, Chatbots and many more. In fact, NLP is indispensable when humans need to interact with computers in natural language. Research in NLP has evolved from the era of punch cards and batch processing to the era of Google [34]. From the time when internet came into existence to the time when social networks (such as Facebook, LinkedIn etc.) were evolved, there was very less amount of data available on the internet. But, presently there is a huge amount of data available over the internet which is also highly unstructured. To deal with this unstructured data, NLP comes to rescue. NLP manages this data for us according to our requirement. One example of such arrangement is Google search engine. We know that a lot of data is present on internet and Google search engine sorts the data using key words related to our requirement. There are many more NLP applications that help us in our daily lives such as Siri [35], Google Help Assistant [36] etc.

### 1.3 Eastern perspectives of Intelligence

Several theories have been put forward for describing and understanding human intelligence. In western science, many theories are based on information processing perspective. The basic idea in these is that intelligence is a product of information processing capabilities of the brain and the mind is an emergent property of the brain too. There are several issues that are debatable with this position and the debate continues. However, the basic outcome of this viewpoint is that the algorithms described above focus only on processing information and training an AI system for a particular task without any regard to the big picture or its utility as a piece in a bigger application and interoperability of the systems so designed. The Values perspective is somewhat understated in these systems.

In the Eastern Philosophy, human intelligence has four different aspects: mann, buddhi, ahamkara, and chitta.In Kathopanishad [37] perspective**,** the body is the chariot, senses are the horses, ‘mann’ (mind) constitutes the reins, ‘buddhi’ (brain) is the charioteer. If mind follows the dictates of ‘buddhi’ it will be safe otherwise senses will run amok. This is a more comprehensive model and provides a window into the workings of the human intellect. More importantly, it provides a good framework for designing AI that has both the quality perspective and the Values perspective rather than focusing only on the quality perspective in a very narrow interpretation of the term quality. This discussion leads to the AI Value Alignment Problem that is described in section 2.

| No. | Name | Functions | Nature | Presiding deity |
|---|---|---|---|---|
| **1.** | Mann | Receives stimuli through organs of perception from external sources | Indecision or doubt | Moon |
| **2.** | Buddhi | Analyses situations or stimuli received and determines cognition faculty | Decision making | Brahma |
| **3.** | Ahamkara | Sense of doership and enjoyership | Ego | Rudra |
| **4.** | Chitta | Recollection of past experiences or events | Storehouse ormemory | Vasudeva |

## 1. The Value Alignment Problem (VAP)

Progress in creating AI systems has thrown up some questions that need to be addressed very urgently for AI to be acceptable. What is acceptable AI or what is expected from AI? One view is that it must have both Quality and Values. Measure of quality depends on the paradigm that is employed for judging the utility of the system and typically it is the market. Quality is local to the problem being solved and typically gets reflected in the utility function that the AI system attempts to optimize. An AI system can be deemed to have Quality if it effectively solves the problem or performs the task for which it was designed. An AI system can be made acceptable by adding values to it so as to ensure that it acts according to the accepted human values. This is typically referred to as the Value Alignment Problem (VAP).

The concept of human values is an overarching global concept. A typical solution may appear to solve a given problem and yet may not pass the value test. As long as system is playing chess or Go, the rules of the game are the only Values it needs to have. But, in the case of a self-driving vehicle, only the rules for driving are not enough. Along with the driving rules, AI system should be fed with values to make an acceptable autonomous AI system. In the absence of values, decisions taken by the AI system may not be acceptable. The fundamental issue is to decide what values to incorporate in an AI system and how to incorporate them.

Values can be defined in different ways from religious to philosophical point of view of what is meant by a good action [38]. The research area of machine ethics has recently emerged as a subfield of AI focusing on the task of ensuring ethical behavior of artificial agents [39]. “One of the issues off course, is to define what exactly these values are, because people might have different cultures, come

from different parts of the world, have different socio-economic backgrounds", says Stefano Ermon. Roman Yampolskiy states, "it is very difficult to encode human values in a programming language, and we as humanity do not agree on common values, and even parts we do agree on changes with time" [40].

Ethics of the AI system reflects the ethics of the designer/user. Ethics enable a final choice to be made among contending alternatives. It is not always possible to explain why a particular alternative has been chosen by the AI system. If any system is to gain significant control over the future, then it is imperative that the system be constructed to act according to the interests of not only its operators, but all humanity, and perhaps all sapient beings [41].

## 2. Extant Approaches for solving the AI VAP

The sudden surge in research papers in 2019 stands as a testimony to the fact that the AI VAP is receiving considerable attention. Some recent attempts on how AI should handle ethical issues are as follows. Bremner et al. [42] describe a layered architecture in which a separate layer is used to handle the ethical reasoning, ethically moderating the robot actions. An attempt at designing an elder care robot was made by Anderson et al. [43] where a set of predefined actions are guided ethically in order of priority. Adamson et al. [44] present an overview of IEEE's current activities related to AI ethics. They argue that human values must drive our future autonomous systems in a way that both protects and benefits humanity. Robertson et al. [45] explain a method for embedding ethics into the design and use of an endoscopic capsule for diagnosis and drug delivery, using a code design approach to reduce the end user risk.Some of the extant approaches for solving AI VAP are as follows.

### 3.1 Using moral stories

A wealth of data is produced in many cultures in the form of written stories. Stories contain socio-cultural knowledge, social protocols, commonly shared knowledge etc. Values canbe learned through these stories. Riedl and Harrison stated that an AI system can learn human values by reading stories [46]. They used RL to teach values to AI using stories. A reward signal is given to the AI agent to encourage it to solve a given problem and discourage it from performing harmful actions. A value aligned reward signal rewards the AI agent for doing what humans would do in same situation and penalizes the agent if it performs actions otherwise. This method faces two challenges while producing

a reward signal from a plot graph. First, it may be possible to execute any of the trajectory tree successors immediately in the current world state. Second, there is possibility of plot graph containing plot points that are not executable by the AI agent. Also, the relationship between the reward function and behavior is complex. The main limitation is that AGI is not addressed by their work. The technique of learning reward signal is best for the AI agents who have limited set of goals but need to interact with humans to achieve their goals.

Scheharzade [47], proposed by D.K. Elson, is an automated story generator that attempts to tell a story about any topic requested by a human. Scheharzade does not rely on hand-built knowledge for telling a story. If it does not have a model for the topic asked, it asks people on the internet (via Amazon's Mechanical Turk Service) to write the example stories in natural language and learns a new model from the example stories. The Scheharzade system has the ability to learn a procedure of how a story on some unknown topic can be presented without any prior knowledge.

Some more approaches mentioned by Nick Bostrom [48]are as follows.

### 3.2 Explicit representation

In this approach, code for all the values that are to be inculcated in AI is written i.e. all the desirable values are hand coded. However, transferring all the desired human values into AI through writing a code in computer language is not cumbersome, to say the least.

### 3.3 Evolutionary selection

Humans are believed to have evolved with the set of values that are desirable to their livelihood. This provides motivation for use of evolutionary methods to solve AI VAP.

### 3.4 Associative value accretion

Humans acquire values through interaction with their environment. An AI system is built in this approach on the same principle i.e. by enabling the AI system to acquire values as it interacts with the environment.

### 3.5 Motivational scaffolding

This approach states a method in which initially small goals are given to the AI with relatively simple final goals. In this approach, the seed AI is given simple goalsand these goals are replaced with more complex ones. Thus, it would develop into a sophisticated representational structure.

### 3.6 Value learning

Another approach to this problem is Value Learning. This approach involves using the AI's intelligence to learn the values that are to be inculcated in it. To do this, AI must be provided with some criteria or method of choosing a set of desirable values. Thus, an AI could be built that could act according to the predefined set of values. Such AI would refine itself continually as it learns more about the world.

### 3.7 Institution design

Any task that we perform is not only motivated by our society but also how this society is organized. For instance, we live in a democratic country, therefore our will of completing a task might be a composite or average of the wills of the wills that lives in our democratic society, whereas, a person or a group that is organized in strong dictatorship society might act as if it had a will of performing a task that is similar to the will of dictator. Therefore, it might be concluded that motivation of the AI for performing a task not only depends on the motivation of its constituent subagents but also on how these subagents are organized.Thus, an attempt could be made to shape motivation of the AI by designing appropriate institutions for it.

## 4 Value based education

The VAP in AI agents is so big that it has routes in all the fields. All the AI techniques described above could be needed to solve this problem with each technique playing its own role in solving this problem. Deep learning can simulate 'mann' (mind) of the agent by training its data for a particular problem. But the mann alone would not be able to understand whether the task to be performed is good or not (i.e., ethical or not). For this, agent will require buddhi along with values. The task then is to teach values to the agent. And this task could be completed with the help of RL / DRL where the objective of the object will be to achieve its goal while neglecting the obstacles (like 'mohh' and 'maya') in its path.

Turing [49] proposed that instead of trying to simulate the adult human mind through programming, we should try to produce a program which simulates a child's mind. He divided this problem further into two parts i.e. the program which simulates child mind and the education process. And these two remain very close. The motivation behind this approach is that a child learns everything from initial

stage by observing and experiencing the things together with the education from where it learns the basic difference between the good or bad.

Education plays an important role in instilling values into a human. Therefore, appropriate training could instill values in AI as well. Education systems that can be studies to draw inspiration for this purpose include the Gurukul System [50], the Shantiniketan System [51, 52] etc.

### 4.1 DEI Education policy

The DEI Education Policy(DEP) [53] was conceived by His Holiness DrMB LalSahab, the Founder Director of DEI in 1975. The policy strives towards the fulfillment of the Holy Mandate given by His Holiness Sir ShabajiMaharaj, the Founder of Dayalbagh "Education, more Education, Education made perfect is the only panacea for all our country's ills and evils". The DEP has been implemented right from the inception of DEI University in 1981 and has withstood the test of time. It has been widely appreciated. The mission objective is to evolve a well-rounded complete person.

Complete person means a person who possesses the qualities of the four Varnas- qualities of a true Brahmin, qualities of a true Kshatriya, qualities of a true Vaishya, and quality of a true Shudra. The true Brahmin is one who is familiar with Vidya (knowledge) of Brahma (God) and imparts his knowledge to all seekers without any recompense and feels happy on doing so. The true Kshatriya is one who has a strong body and brave heart and is ready to sacrifice his life in protecting the others. The true Vaishya is one who may develop industries and handicrafts and acquire wealth and be ready to spend the wealth, without any selfishness, on things of public good. The true Shudra is the one who believes in performing the 'Seva' (service).

The DEP system for evolving a Complete Person consists of components that inculcate three kinds of values – academic values, moral and spiritual values and social sensibilities. There is a fourth subsystem, which maintains its quality and enhances it continually.

Academic objectives consist of elements like Intellectual Strength, Scientific Temper, Self-reliance, interdisciplinary outlook, acquisition of aptitude, independent thinking, reasoning ability and habit of learning. Similarly, moral and spiritual values would consist of emotional maturity, ethical values, simple living, selfless service, humility, truthfulness, dignity of labour and high moral character. Social sensibilities include national character and heritage, tolerance for diversity, national integration,

understanding of rural life, civic sense, respect for rights of others, and brotherhood of man. Quality stands by itself.

There is a second subsystem, namely educational and curricular features, which consist of Intellectual Activity, Social Activity, Physical Activity. The curriculum is broad based and interdisciplinary in its outlook and approach. The core courses for all UG students include several aspects like Indian culture, Comparative Study of Religion, General Knowledge, Scientific Methodology, Work Experience Courses and Co-curricular activities.

Infrastructure for learning is provided. Student welfare measures are provided in terms of free or affordable education and assistance through means based. There is a provision of attachment of a University to below University institutions. Special efforts are available for weaker sections and also for gifted students.

## 5 Challenges in tackling AI VAP

The literature survey presented in the preceding sections clearly brings out the fact that the AI VAP is an important problem that requires urgent solution. The VAP is not to be solved only for Superintelligent machines as and when they are produced. It is important for all kinds of autonomous and even semi-autonomous systems that have any decision-making capabilities. Every time a decision is made by an AI entity it must be aligned to the human value system. This is referred to as Acceptable AI.

The problem of creating an Acceptable AI throws up certain important challenges as follows.

(i) How does one specify the Value System with which the behavior of the autonomous systems is to be aligned? This is particularly difficult because Values are construed differently in different societies and at different times in the same society. What was unacceptable in the last century may become acceptable in this century.

(ii) New problems keep coming up. What was not even contemplated about in the last century may become an important consideration in this one. The Trolley Problem that has come into prominence with the development of autonomous cars is a case in point.

(iii) Can the specification of the Value system be broken down into smaller components each addressing a particular value that would be easier to inculcate?

(iv) What would be these individual components or values that must be included in an Acceptable AI?

(v) Can some specification of Values be developed such that they are classified into categories as follow?

a. Hard constraints wherein it is imperative that the system adhere to the value concerned and no violation is acceptable.

b. Soft Constraints wherein it is desirable that the system adhere to the value concerned but only when it does not have to violate a Value in class a above.

(vi) How does one then ensure that the system that is designed with these individual values one at a time is then consistent in the sense that it does not arrive at disparate actions according to different values making a final decision of a particular action complicated?

(vii) Can one create a hierarchical system that filters out the unacceptable courses of action when they do not satisfy the Value alignment requirement leaving only those actions that pass these tests so that the design criteria can then be utilized to select the best out of the remaining options that all meet the Value alignment criteria?

(viii) How does one design a system that is aligned with the Value system specified while addressing all the above questions satisfactorily?

(ix) Even if a system is designed with the considerations listed above how does one ensure that the system does not violate any consideration even when the costs involved of adhering to the constraints are higher?

(x) Even if it is designed to accept the higher costs and still adhere to the Value system, how does one show that this is indeed the case in all possible situations without fail? This is with particular reference to the requirement of AI being Explainable so that the user can see the reasoning chain and be convinced that the decision taken by the autonomous system is the best possible according to the conditions prevalent at the particular time.

Although attempts have been made to address these questions in the literature there are several gaps that still need to be addressed. This thesis is an attempt to fill some of these gaps.

## 6 Proposed work

The overarching goal of this work is the development of a framework for addressing the AI VAP. Specifically, it is proposed to make attempts in the following directions.

(i) The first step is to identify the set of Values that are deemed to be necessary to be inculcated in the AI system. A deep study of DEP Policy and its elements along with its implementation shall be made to converge on to a set of Values that will have to be provided in the AI systems.

(ii) Moral stories and religious text shall be analyzed using Natural Language Processing techniques to infer the essence of these documents. The knowledge extracted from these documents will be utilized to create a knowledge base with a suitable inference engine for enabling the test of a given hypothesis for value alignment.

An attempt can be made to prepare a schema to represent different entities related to schema. This representation schema will consists of ethically relevant **features,** their incumbent **duties** to either maximize or minimize the features, **actions** represented as a tuple of integer degrees of presence and absence of the ethically relevant features, **cases**- represented as a tuple of the differences of the corresponding duty satisfaction/ violation degrees of the two actions and the **principle** of ethical action preference, i.e. principle which will determine currently ethically preferable actions [54].

The entities will be represented to find out features, duties, actions, cases and principles relevant to context at hand. After identifying these five elements, these elements will be stored in proper data structures with corresponding algorithmic implementations. Case-based learning will be employed to train the system.

(iii) The knowledge base created in step (ii) above would provide filters. The proposed actions that the system arrives at would be checked by these filters. Even if the system arrives at a particular action that it deems to be the most optimal given the current situation of the parameters at hand, it would not be implemented unless it is passed by the filters. This is

akin to the concept of robust optimization. The idea of robust optimization is that it may be preferable to implement a solution of lesser fitness provided it is on plateau when fitness function is plotted in parameter space. Maintaining the system parameters at precisely the optimum value could be a challenge and any deviation, however slight, would affect the fitness drastically in this case. In the case the solution is on a plateau in the parameter space, parameter changes would not impact the fitness. Similarly, it would be necessary to accept a solution with lesser fitness in the light of satisfying value filters. An architecture consisting of different layers could be designed as follows.

a). Goal Layer: Propose goal to be accomplished.

b). Task Layer: Generate a set of tasks that might fulfill the current goal proposed by the goal layer.

c). Action Layer: Generate the actions needed for task execution.

d). Ethical Layer: Takes as input the tasks that the task layer suggests and outputs the most appropriate task to the action layer.

(iv) Learning by examples has also been shown to be a very powerful mechanism for implementing learning algorithms for teaching values. An attempt would be made to generate example set for the Values identified and train the system using these examples. The outputs of this system would be validated through the knowledge base as specified above.

A framework can be designed for which data will be gathered and abstracted into a perception list (list containing Boolean values representing the current state of the environment). Then, these values will be used to determine the duty satisfaction/violation values (at what extent the value is satisfied or violated) via a decision tree learned from examples, for each duty and each possible action, represented as an action matrix. The action matrix then, will be sorted from the most ethically preferable to the least ethically preferable. The most ethically preferable action is then communicated to the agent which then performs it.

(v) An important aspect in creating an acceptable AI is for it to provide a transparent explanation of its reasoning process. An interface with the user would be created in order to explain the underlying process. This would be facilitated by the presence of the knowledge base.

**References**


**1.** D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," Nature, vol. 529, no. 7587, pp. 484, 2016.

**2.** D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., "Mastering the Game of Go without Human Knowledge," Nature, vol. 550, no. 7676, pp. 354, 2017.

**3.** M. A. Hawas, "Are We Intentionally Limiting Urban Planning an Intelligence? A Causal Evaluative Review and Methodical Redirection for Intelligence Systems," IEEE Access, vol. 5, pp. 13253–13259, 2017.

**4.** M. Campbell, A. J. HoaneJr, and F.-h. Hsu, "Deep Blue," Artificial intelligence, vol. 134, no. 1-2, pp. 57–83, 2002.

**5.** L. S. Gottfredson, "Mainstream Science on Intelligence: An Editorial with 52 Signatories, History, and Bibliography," 1997.

**6.** N. Bostrom, "How long before superintelligence,?" International Journal of Futures Studies, vol. 2, 1998.

**7.** Medium. (2019). Why AGI is Achievable in Five Years. [online] Available at: https://medium.com/intuitionmachine/near-term-agi-should-be-considered-as-a-possibility-9bcf276f9b16.

**8.** F.Rothganger, "The Dawn of the Real Thinking Machine," IEEE Spectrum, vol. 54, pp. 22, June 2017.

**9.** J. Hawkins, "What Intellingent Machines need to learn from the Neocortex," IEEE Spectrum, vol. 54, pp. 36, June 2017.

**10.** E. Strickland, " From Animal Intelligence to Artificial Intelligence," IEEE Spectrum, vol. 54, pp. 42, June 2017.

**11.** L. Gomes, "The Neuromorphic Chip's Make-or-Break Moment," IEEE Spectrum, vol. 54, pp. 52, June 2017.

**12.** L. Gomes, "The Neuromorphic Chip's Make-or-Break Moment," IEEE Spectrum, vol. 54, pp. 54, June 2017.

**13.** J. kumagai, "Navigate like a Rat", IEEE Spectrum, vol. 54, pp. 60, June 2017.

**14.** C. Koch and G. Tononi, "Can we Quantify machine Conciousness," IEEE Spectrum, vol. 54, pp. 66, June 2017.

**15.** S. L. Poczter and L. M. Jankovic, "The Google Car: Driving Toward A Better Future?," Journal of Business Case Studies (Online), vol. 10, no. 1, pp. 7, 2014.

**16.** K. Bimbraw, "Autonomous Cars: Past, Present And Future, A Review Of The Developments In The Last Century, The Present Scenario And The Expected Future Of Autonomous Vehicle Technology," in 2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO), vol. 1, pp. 191–198, IEEE, 2015.

**17.** M. Campbell, M. Egerstedt, J. P. How, and R. M. Murray, "Autonomous Driving In Urban Environments: Approaches, Lessons And Challenges," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 368, no. 1928, pp. 4649– 4672, 2010.P.

**18.** R. Moha, P. K. Sree, "An Extensive Survey on Deep Learning Applications," IJIACS, vol. 6, 2017.

**19.** M. N. Sadiku, M. Tembely, and S. M. Musa, "Deep learning," International Research Journal of Advanced Engineering and Science, vol. 2, pp. 77–78, 2017.

**20.** Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436, 2015.

**21.** Z. Wang, "The Applications of Deep Learning on Traffic Identification," BlackHat USA, pp. 21– 26, 2015.

**22.** A. Hammoudeh, A Concise Introduction to Reinforcement Learning.

Doi: 10.13140/RG.2.2.31027.53285,Feb2018.

**23.** M. A. Mueller, "Reinforcement Learning: MDP applied to autonomous navigation," Machine Learning and Applications: An International Journal, vol. 4, pp. 01–10, Dec 2017.

**24.** L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A survey," Journal of Artificial Intelligence Research, vol. 4, pp. 237–285, 1996.

**25.** R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction.MIT press, 2011.

**26.** H. Van Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-Learning," In Thirtieth AAAI Conference on Artificial Intelligence, 2016.

**27.** V. Francois-Lavet, P. Henderson, R. Islam, M. G. Bellemare, J. Pineau, et al., "An Introduction to Deep Reinforcement Learning," Foundations and Trends in Machine Learning, vol. 11, no. 3-4, pp. 219–354, 2018.

**28.** David Silver, Deep reinforcement learning, a tutorial at ICML 2016. http://icml.cc/2016/tutorials/deep_rl_tutorial.pdf.

**29.** T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous Control with Deep Reinforcement Learning," arXiv preprint arXiv:1509.02971, 2015.

**30.** D. G. Wilson, S. Cussat-Blanc, H. Luga, and J. F. Miller, "Evolving Simple Programs for Playing Atari Games," in Proceedings of the Genetic and Evolutionary Computation Conference, pp. 229–236, ACM, 2018.

**31.** M. Hausknecht, J. Lehman, R. Miikkulainen, and P. Stone, "A Neuroevolution Approach to General Atari Game Playing," IEEE Transactions on Computational Intelligence and AI in Games, vol. 6, no. 4, pp. 355–366, 2014.

**32.** E. Zitzler, K. Deb, and L. Thiele, "Comparison of Multiobjective Evolutionary Algorithms: Empirical Results," Evolutionary computation, vol. 8, no. 2, pp. 173–195, 2000.

**33.** H. Chung, J. Park, and S. Lee, "Digital Forensic Approaches for Amazon Alexa Ecosystem," Digital Investigation, vol. 22, pp. S15–S25, 2017.

**34.** E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research," IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48–57, 2014.

**35.** M. Assefi, G. Liu, M. P. Wittie, and C. Izurieta, "An Experimental Evaluation of Apple Siri and Goo6gle Speech Recognition," Proccedings of the 2015 ISCA SEDE, 2015.

**36.** https://assistant.google.com/

**37.** S. Chinmayananda, "Kathopanishad,"Chinmaya Mission, 2006.

**38.** S. O. Ogunlere and A. O. Adebayo, Ethical issues in computing sciences. Jan 2015.

**39.** C. Shulman, H. Jonsson, and N. Tarleton, "Machine Ethics andSuperintelligence," Reynolds and Cassinelli, pp. 95–97, 2009.

**40.** https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/?cn-reloaded=1

**41.** N. Soares, "The Value Learning Problem," Machine Intelligence Research Institute, Berkley, 2015.

**42.** Bremner, L. A. Dennis, M. Fisher, and A. F. Winfield, "On proactive, transparent, and verifiable ethical reasoning for robots," Proceedings of the IEEE, vol. 107, no. 3, pp. 541–561, 2019.

**43.** M. Anderson, S. L. Anderson, and V. Berenz, "A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior paradigm," Proceedings of the IEEE, vol. 107, no. 3, pp. 526–540, 2018.

**44.** G. Adamson, J. C. Havens, and R. Chatila, "Designing a value-driven future for ethical autonomous and intelligent systems," Proceedings of the IEEE, vol. 107, no. 3, pp. 518–525, 2019.

**45.** L. J. Robertson, R. Abbas, G. Alici, A. Munoz, and K. Michael, "Engineering based design methodology for embedding ethics in autonomous robots," Proceedings of the IEEE, vol. 107, no. 3, pp. 582–599, 2019.

**46.** M. O. Riedl and B. Harrison, "Using Stories To Teach Human Values To Artificial Agents," In Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, 2016.

**47.** D. K. Elson, "Modeling Narrative Discourse," PhD thesis, Columbia University, 2012.

**48.** N. Bostrom, "Value Acquisition," in Superintelligence: Paths, Dangers, Strategies, Oxford, United Kingdom: Oxford University Press, 2016, pp. 226–255.

**49.** A. Turing, “Computing Machinery and Intelligence,” Mind, vol. 59, no. 236, p. 433, 1950.

**50.** A. N. Basu, “Education In Modern India,” Calcutta: Orient Book Co, 1947.

**51.** J. Mondal, “Tagore’s Education System in Shantiniketan: A Geographical Approach for Survival of Mankind,” Imperial Journal of Interdisciplinary Research, vol. 3, 2017.

**52.** T. Pushpanathan, “Rabindranath Tagore’s Philosophy of Education and Its Influence on Indian Education,” International Journal Of Current Research And Academic Review, vol. 1, no. 4, pp. 42–45, 2013.

**53.** “Progress of Education in Dayalbagh,” Dayalbagh Educational Institute, Dayalbagh, Agra, First Edition, Feb 2005.